A NONLINEAR OBSERVATION MODEL FOR REMOVING NOISE FROM CORRUPTED SPEECH LOG MEL-SPECTRAL ENERGIES

Jasha Droppo, Alex Acero, and Li Deng

Microsoft Research, One Microsoft Way, Redmond, Washington, USA

ABSTRACT

In this paper we present a new statistical model, which describes the corruption to speech recognition Mel-frequency spectral features caused by additive noise. This model explicitly represents the effect of unknown phase together with the unobserved clean speech and noise as three hidden variables. We use this model to produce noise robust features for automatic speech recognition.

The model is constructed in the log Mel-frequency feature domain. In addition to being linearly related to MFCC recognition parameters, we gain the advantage of low dimensionality and independence of the corruption across feature dimensions.

We illustrate the surprising result that, even when the true noise Mel-frequency spectral feature is known, the traditional spectral subtraction formula is flawed. We show the new model can be used to derive a spectral subtraction formula which produces superior error rate results, and is less sensitive to tuning parameters.

Finally, we present results demonstrating that the new model is more general than spectral subtraction, and can take advantage of a prior noise estimate to produce robust features, rather than relying on point estimates of noise.

1. INTRODUCTION

Automatic speech recognition systems without explicit provisions for noise robustness degrade quickly in the presence of additive noise. As a consequence, how to best add noise robustness to such systems is an area of active research.

This paper is organized as follows. Section 2 illuminates a flaw with standard spectral subtraction. One would expect that when the true noise spectra are available, basic spectral subtraction should do a good job of removing the noise. In Section 3, we propose a new non-linear model for the environment, and derive a better spectral subtraction formula. Section 4 presents how this non-linear model can be embedded in a Bayesian framework. The model is leveraged to produce posterior probabilities about the unobserved clean speech, and consequently, noise robust features. Careful attention is paid to the accuracy of the model, and initial experiments show that it is useful for producing noise robust features.

All experiments were conducted using the data, code, and training scripts provided within the Aurora 2 evaluation framework[1]. The task consists of recognizing strings of English digits embedded in a range of artificial noise conditions. Although the framework provides for evaluation against three sets of data, we present here results for set A only. The acoustic model used is the "clean" acoustic model trained with the standard scripts on uncorrupted data. It consists of eleven whole word models, containing a total of 546 diagonal 39 dimensional Gaussian mixture components.

To conform with our probabilistic models, the feature generation was modified slightly from the reference "FE V2.0" implementation. In particular, we replaced the log energy feature with c_0, and changed from using spectral magnitude to using power spectral density as the input to the Mel-frequency filterbank. Additionally, the decoder was modified to use uncertainty estimates from the noise removal process to better decode the speech as in [2].

2. REVIEW OF SPECTRAL SUBTRACTION

Standard spectral subtraction is motivated by the observation that noise and speech spectra mix linearly, and therefore their log spectra should mix according to

$$|Y[k]|^2 = |X[k]|^2 + |N[k]|^2$$

Typically, this equation is solved for $|X[k]|^2$, and a maximum attenuation floor F is introduced to avoid producing negative power spectral densities.

$$\left|\hat{X}[k]\right|^2 = |Y[k]|^2 \max\left(\frac{|Y[k]|^2 - |N[k]|^2}{|Y[k]|^2}, F\right) \quad (1)$$

We ran several experiments to examine the performance of Eq. 1 using the true spectra of n, and floors F from e^{-20} to e^{-2}. The true noise spectra were computed from the true additive noise time series for each utterance. The first row of Table 1 reports the digit error rates.

Table 1. Average digit error rates for set A of the Aurora 2 task. Two versions of a log-domain spectral subtraction algorithm with true noise feature estimates are compared.

	Floor				
Method	e^{-20}	e^{-10}	e^{-5}	e^{-3}	e^{-2}
Standard (Eq. 1)	87.50	56.00	34.54	11.31	15.56
Proposed (Eq. 9)	6.43	5.74	4.10	7.82	10.00

It is somewhat surprising that *even when we know the noise spectra exactly*, spectral subtraction can not do a perfect job. The next section is dedicated to building a better model for the environment, which can be used to derive better noise removal formulae.

3. A NEW MODEL OF THE ENVIRONMENT

The front end processes each spectral frame by passing it through a magnitude-squared operation, a Mel-frequency filterbank, and a logarithm.

For our case where the noisy signal is a linear combination of speech and noise, $Y[k] = X[k]+N[k]$, the noisy log Mel-spectral features y_i can be directly related to the unobserved spectra $X[k]$ and $N[k]$.

$$\begin{aligned}\exp y_i &= \sum_k w_k^i \left|X[k]\right|^2 + \sum_k w_k^i \left|N[k]\right|^2 \\ &+ \sum_k w_k^i \left|X[k]\right| \left|N[k]\right| \cos\theta_k \qquad (2)\end{aligned}$$

Here, w_k^i is the kth coefficient in the ith Mel-frequency filterbank. The variable θ_k is the phase difference between $X[k]$ and $N[k]$. When the clean signal and noise are uncorrelated, the θ_k are uncorrelated and have a uniform distribution over the range $[-\pi, \pi]$.

Eq. 2 can be re-written to show how the noisy log spectral energies y_i are a function of the unobserved log spectral energies x_i and n_i.

$$\begin{aligned}\exp y_i &= \exp x_i + \exp n_i + 2\alpha_i \exp \frac{x_i + n_i}{2} \qquad (3)\\ \alpha_i &= \frac{\sum_k w_k^i \left|X[k]\right| \left|N[k]\right| \cos\theta_k}{\sqrt{\sum_k w_k^i \left|X[k]\right|^2}\sqrt{\sum_k w_k^i \left|N[k]\right|^2}}\end{aligned}$$

As a consequence of this model, when we observe y_i there are actually *three* unobserved random variables. The first two are obvious: the clean log spectral energy and the noise log spectral energy that would have been produced in the absence of mixing. The third variable, α_i, accounts for the unknown phase between the two sources.

If the magnitude spectra are assumed constant over the bandwidth of a particular filterbank, the definition of α_i collapses to a weighted sum of several independent random variables:

$$\alpha_i \approx \sum_k \frac{w_k^i}{\sum_j w_j^i} \cos\theta_k. \qquad (4)$$

Figure 1d shows the true distributions of α for a range of frequency bins. They were estimated from a set of joint noise, clean speech, and noisy speech data by solving for the unknown α. The higher frequency, higher bandwidth filters produce α distributions that are more nearly Gaussian. As the bandwidth increases, so does the number of effective terms in Eq. 4, and the central limit theorem begins to apply. In practice, a frequency-dependent Gaussian approximation $p_{\alpha_i}(\alpha_i) = N(\alpha_i; 0, \sigma_{\alpha_i}^2)$ works well.

3.1. Conditional observation probability

Eq. 3 places a hard constraint on the four random variables, in effect giving us three degrees of freedom. We can express this by solving for y and writing the conditional probability distribution,

$$p(y|x, n, \alpha) = \delta\left(y - \ln\left(e^x + e^n + 2\alpha e^{\frac{x+n}{2}}\right)\right). \qquad (5)$$

The conditional probability $p(y|x, n)$ is found by forming the distribution $p(y, \alpha|x, n)$ and marginalizing over α. Note that here we are assuming $p(\alpha|x, n) = p(\alpha)$, which is reasonable.

$$\begin{aligned}p(y|x, n) &= \int_{-\infty}^{\infty} p(y|x, n, \alpha) p_\alpha(\alpha) d\alpha \\ &= \int_{-\infty}^{\infty} \delta\left(y - \ln\left(e^x + e^n + 2\alpha e^{\frac{x+n}{2}}\right)\right) p_\alpha(\alpha) d\alpha\end{aligned}$$

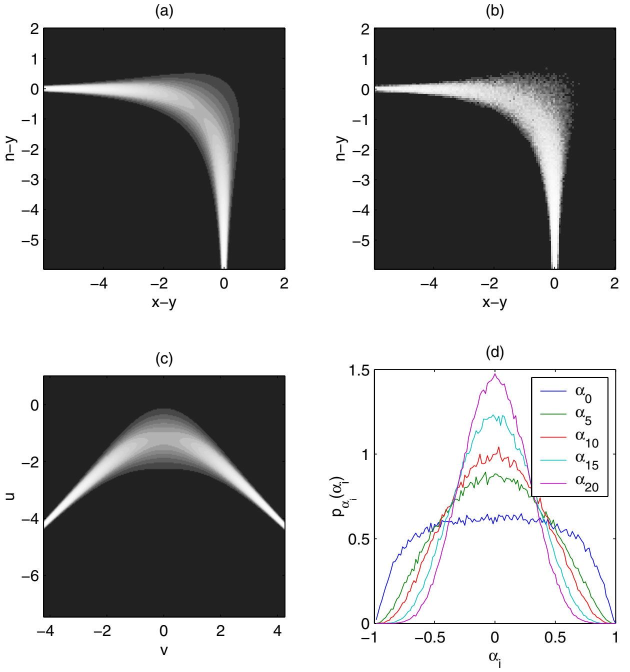

Fig. 1. (a) Conditional observation probability $p(y|x, n)$ with Normal approximation for $p_\alpha(\alpha)$. (b) Sample distribution for filterbank 13 in set A, noise 1, SNR 10. (c) Normal approximation of $p(y|x, n)$ as a function of v. (d) True distribution of α for several frequency bins.

Now, we can use the identity

$$\int_{-\infty}^{\infty} \delta\left(f(\alpha)\right) p_\alpha(\alpha) d\alpha = \sum_{\{\alpha: f(\alpha)=0\}} \frac{p_\alpha(\alpha)}{\left|\frac{d}{d\alpha} f(\alpha)\right|}$$

to evaluate the integral in closed form:

$$p(y|x, n) = \frac{1}{2} \exp\left(y - \frac{x+n}{2}\right) p_\alpha\left(\frac{e^y - e^x - e^n}{2e^{\frac{x+n}{2}}}\right) \qquad (6)$$

When we introduce the approximation $p_\alpha(\alpha) = N(\alpha; 0, \sigma_\alpha^2)$, the likelihood function becomes

$$\ln p(y|x, n) = y - \frac{x+n}{2} - \frac{1}{2}\ln 8\pi\sigma_\alpha^2 - \frac{\left(e^y - e^x - e^n\right)^2}{8\sigma_\alpha^2 e^{(x+n)}}. \qquad (7)$$

Figure 1a contains a plot of this conditional probability distribution. Note that due to the shift invariance of the model, there are only two independent terms in the plot. Figure 1b shows an equivalent plot, directly estimated from data, confirming Eq. 7.

3.2. Relationship to Spectral Subtraction

We can use Eq. 7 to derive a new formula for spectral subtraction. The first step is to hold n and y fixed, and find a maximum likelihood estimate for x. Taking the derivative with respect to x in Eq. 7 and equating it to zero results in

$$e^{x-n} = \sqrt{\left(e^{y-n} - 1\right)^2 + \left(2\sigma_\alpha^2\right)^2} - 2\sigma_\alpha^2. \qquad (8)$$

This formula is already more well-behaved than standard spectral subtraction. The first term is always real because we are taking

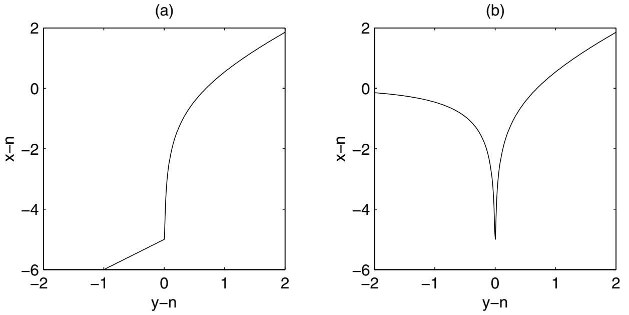

Fig. 2. (a) Output SNR versus input SNR for spectral subtraction (Eq. 1). (b) The new model (Eq. 9) treats the points $y < n$ differently.

the square root of the sum of two positive numbers. Furthermore, the magnitude of the second term is never larger than the magnitude of the first term, so both sides of Eq. 8 are non-negative. The entire formula has exactly one zero, at $n = y$. This automatically prevents us from taking the logarithm of any negative numbers during spectral subtraction, allowing us to severely relax the maximum attenuation floor F.

When we set $\sigma_\alpha^2 = 0$ and solve for x in Eq. 8, the result is the familiar spectral subtraction equation, with an unexpected absolute value operation.

$$\hat{x} \quad = \quad y + \ln\left|1 - e^{n-y}\right| \tag{9}$$

The difference between Eq. 1 and Eq. 9 is confined to the region $y < n$, as shown in Figure 2. Eq. 9 has more reasonable behavior in this region. As the observation becomes much lower than the noise estimate, the function approaches $x = n$. Our model indicates the most likely state is that x and n have similar magnitudes and are experiencing destructive phase interference.

Table 1 compares the relative accuracy of using Equations 1 and 9 for speech recognition, when the true noise spectra are available. Although our new method does not require a floor to prevent taking the logarithm of a negative number, we include it in our results because it does yield a small improvement in error rate.

Regardless of the value chosen for the floor, the new method outperforms the old spectral subtraction rule. Although the old method is quite sensitive to the value chosen, the new method is not, producing less than 10% digit error rate for all tests.

4. A BAYESIAN APPROACH

The major advantage of deriving the conditional observation probability, Eq. 7, is that we can embed it into a unified Bayesian model. In this model the observed variable y is related to the hidden variables, including x and n.

$$p(y, x, n) = p(y|x, n)p_x(x)p_n(n)$$

To produce noise-removed features for conventional decoding, we simply take conditional expectations of this model.

$$E[x|y] \quad = \quad \int_{-\infty}^{\infty} x\, p(x|y)\, dx, \text{ where} \tag{10}$$

$$p(x|y) \quad = \quad \frac{\int_{-\infty}^{\infty} p(y, x, n)\, dn}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(y, x, n)\, dx\, dn} \tag{11}$$

The Bayesian approach can additionally produce a variance of its estimate of $E[x|y]$. It has been shown in [2] that this variance can be easily leveraged within the decoder to improve word accuracy. In this uncertain decoding, the static feature stream is replaced with an estimate of $p(y|x)$. The noise removal process outputs high variance for low SNR features, and low variance when the SNR is high. To support this framework, we also need

$$E[x^2|y] \quad = \quad \int_{-\infty}^{\infty} x^2 p(x|y)\, dx, \text{ and} \tag{12}$$

$$p(y) \quad = \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(y, x, n)\, dx\, dn. \tag{13}$$

Better results are achieved with a stronger prior distribution for clean speech, such as a mixture model.

$$p_x(x) = \sum_m p_x(x|m)p_m(m).$$

When a mixture model is used, Equations 10, 11, 12, and 13 are conditioned on the mixture m, evaluated, and then combined in the standard way:

$$\begin{array}{cc} p(y) = \sum_m p(y|m)p(m) & E[x|y] = \sum_m E[x|y, m]p(m|y) \\ p(m|y) = \frac{p(y|m)p(m)}{p(y)} & E[x^2|y] = \sum_m E[x^2|y, m]p(m|y) \end{array}$$

4.1. Approximating the observation likelihood

If the environmental model were linear, the expectations could be directly computed at this point. Unfortunately, the true relationship between y, x, and n, as manifested in the form of $p(y|x, n)$, is quite non-linear. Furthermore, our model is capable of producing joint distributions $p(y, x, n)$ that are not well modeled by a single Gaussian approximation. This complicates obtaining a closed form solution, and necessitates finding a good approximation.

We use a Gaussian approximation along one dimension only, which allows us to preserve the true shape of $p(y|x, n)$, and implement a numerical integration along the remaining dimension.

The weighted Gaussian approximation is found in four steps. We rotate the coordinate space, choose an expansion point, find a second order Taylor series approximation, and express the approximation as the parameters of a weighted Gaussian distribution.

The coordinate rotation is necessary because expanding along x or n directly can be problematic. We choose a 45 degree rotation, which makes $p(y|x, n)$ approximately Gaussian along u for each value of v.

$$u(y, x, n) = \tfrac{1}{\sqrt{2}}\left(x + n - 2y\right), \quad v(y, x, n) = \tfrac{1}{\sqrt{2}}\left(x - n\right) \tag{14}$$

Although the new coordinates u and v are linear functions of y, x, and n, we drop the cumbersome functional notation at this point.

After this change of variables, the conditional observation likelihood becomes,

$$\ln p(y|x, n) \quad = \quad -\frac{1}{\sqrt{2}}u - \frac{1}{2}\ln 8\pi\sigma_\alpha^2 - \frac{\left(1 - e^{\frac{u}{\sqrt{2}}}\left(e^{\frac{v}{\sqrt{2}}} - e^{-\frac{v}{\sqrt{2}}}\right)\right)^2}{8\sigma_\alpha^2 \exp\left(\sqrt{2}u\right)}.$$

The Taylor series expansion point is found by performing our change of variables on Eq. 3, holding v constant, letting $\alpha = 0$, and solving for u. The result is,

$$u_v = v - \sqrt{2}\ln\left(1+\exp\sqrt{2}v\right).$$

The coefficients of the expansion are the derivatives of $p(y|x,n)$ evaluated at u_v.

$$\begin{aligned} p(y|x,n)|_{u=u_v} &= \frac{\ln\left(1+\cosh\sqrt{2}v\right)-\ln 4\pi\sigma_\alpha^2}{2} \\ \left.\frac{d}{du}\ln p(y|x,n)\right|_{u=u_v} &= -\frac{1}{2}\sqrt{2} \\ \left.\frac{d^2}{du^2}\ln p(y|x,n)\right|_{u=u_v} &= -\frac{1+\cosh\sqrt{2}v}{4\sigma_\alpha^2} \end{aligned}$$

The quadratic approximation for $p(y|x,n)$ at each value of v can be expressed as a Gaussian distribution along u. Our final approximation is given by:

$$\begin{aligned} p(y|x,n) &= e^{K_v}N\left(u;\mu_v,\sigma_v^2\right)\text{, where} \qquad (15) \\ \sigma_v^2 &= \frac{4\sigma_\alpha^2}{1+\cosh\sqrt{2}v}, \\ \mu_v &= u_v - \frac{1}{\sqrt{2}}\sigma_v^2\text{, and} \\ K_v &= \frac{1}{2}\ln 2 + \frac{\sigma_\alpha^2}{1+\cosh\sqrt{2}v}. \end{aligned}$$

As Figure 1c illustrates, approximating $p(y|x,n)$ as a Gaussian function of v captures the true shape of $p(y|x,n)$ quite well.

The approximation for $p(y|x,n)$ is complete, and is now combined with the priors $p_x(x)$ and $p_n(n)$ to produce the joint probability distribution. To conform with our approximation of the conditional observation probability, we transform these prior distributions to the (u,v) coordinate space, and write them as a Gaussian in u whose mean and variance are functions of v.

$$p_{x,n}(x,n) = p_x(x)p_n(n) = N(u;\eta_v,\gamma_v^2). \qquad (16)$$

From the joint probability, we compute Equations 10, 11, 12, and 13. Each equation requires at least one double integral over x and n, which is equivalent to a double integral over u and v. For example:

$$\begin{aligned} &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xp(y,x,n)\,dx\,dn \\ = &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\frac{u+v}{\sqrt{2}}+y\right)N(u;\mu_v,\sigma_v^2)N(u;\eta_v,\gamma_v^2)\,du\,dv \\ = &\int_{-\infty}^{\infty} e^{K_v}\left(\frac{\hat{\mu}_v+v}{\sqrt{2}}+y\right)N(\mu_v;\eta_v,\sigma_v^2+\gamma_v^2)\,dv, \\ &\text{where } \hat{\mu}_v = \frac{\sigma_v^2\eta_v+\gamma_v^2\mu_v}{\sigma_v^2+\gamma_v^2}. \end{aligned}$$

Here, we have made use of Eq. 14 for x, as well as Eq. 15 and Eq. 16 for $p(y,x,n)$. The Gaussian approximation enables a symbolic evaluation of the integral over u, but the integral over v remains.

The integration in v is currently implemented as numerical integration, a weighed sum along discrete values of v. For this paper, we use 500 equally spaced points in the range $[-20,20]$. Most of the necessary values can be precomputed and tabulated to speed up computation.

4.2. Experimental results

Just as we did for spectral subtraction, we can test our model by feeding it the true noise log Mel-frequency energies. We then simulate imperfect noise estimates by artificially increasing the variance of these perfect values.

Instead of using an attenuation floor, as we did for spectral subtraction, these experiments leverage the Bayesian framework by introducing a simple diagonal Gaussian mixture prior for clean speech. The number of mixtures is varied from 2 to 8.

Table 2. Average digit error rates for set A of the Aurora 2 task with the Bayesian estimate for clean speech, and uncertainty decoding.

Clean Speech	Artificial Variance		
GMM Mixtures	σ_n^2=0	σ_n^2=0.5	σ_n^2=1.0
2	5.98	10.09	12.10
4	2.63	8.62	11.72
8	2.28	8.04	11.12

The first column of Table 2 shows that our model, together with a simple clean speech prior and known accurate noise estimates, produces very low digit error rates for this task. These error rates are much smaller than those achieved through spectral subtraction. The other columns show that when we simulate imperfect noise estimates, the system degrades gracefully.

5. SUMMARY AND CONCLUSION

One promising architecture for noise removal is to define a probabilistic model that unites the observation together with the unobserved noise and clean speech. Prior information about speech and noise, and possibly its dynamics, can be incorporated with standard inference techniques to obtain estimates of the hidden speech and noise variables.

This paper describes such a system, and tests the performance limits of the architecture. Care is taken such that the conditional observation probability is correct, and is evaluated as accurately as possible.

We show that the new model can be used to derive a superior spectral subtraction rule for the case where point estimates of noise are available, and that it can also take advantage of more general, statistical estimates of the noise input.

6. REFERENCES

[1] H. G. Hirsch and D. Pearce, “The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy condidions,” in *ISCA ITRW ASR2000 ”Automatic Speech Recognition: Challenges for the Next Millennium”*, Paris, France, September 2000.

[2] J. Droppo, A. Acero, and L. Deng, “Uncertainty decoding with SPLICE for noise robust speech recognition,” in *Proc. 2002 ICASSP*, Orlando, Florida, May 2002.